Software work in mathematical biology at Georgia Tech

Maxie Dion Schmidt

maxieds@gmail.com
http://people.math.gatech.edu/~mschmidt34
https://github.com/maxieds

Sandia National Labs
Technical Presentation
Spring 2022

Introduction – Applications of RNA research

- RNA sequencing utilized in cancer research and therapy
- mRNA vaccines are newly available to the public but have been studied for decades
- Allowed for the rapid development of a COVID-19 vaccine
- Research on RNA may eventually play a central role in medical applications

What is mathematical biology?

- Mathematical biology (MathBio) uses mathematical models as theoretical abstractions of the natural structure of living organisms
- In this talk we will discuss my work as a software engineering RA with the *Georgia Tech Discrete Mathematics and Molecular Biology* group (gtDMMB)

RNA basics

RNA basics

- RNA is a single-stranded molecule similar to DNA
- A strand of RNA has a backbone of alternating sugar (ribose) and phosphate groups
- Each sugar has one of four base types attached to it (**A–U–C–G**)
- Each of the **A–U–C–G** bases can fold to form bonds in pairs

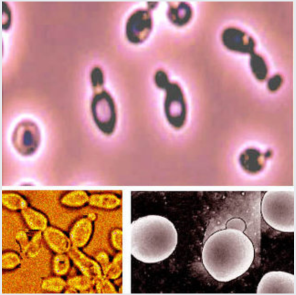
Arc diagrams – Discussion example – S. Cerevisiae (yeast)

GGUUGCGGCCAUAUCUACCAGAAAGCACCGUUUCCCGUCCGAUCAACUGUGUUAAGCUGGUAGA

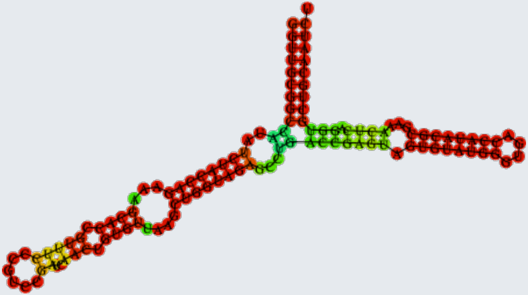
(((((((((....((((((((...((((.(((((......))..))).))))....))))))))

GCCUGACCGAGUAGUGUAUGGGUGACCAUACGCGAAACUCAGGUGCUGCAAUCU

.....(((((((.((((((((....))))))))...)))).))))))))))))).



(Actual microscopic views)



(Radial view of 2D MFE structure)

RNA secondary structures

- RNA base sequence and can have more than one 2D or 3D structure
- Obtaining the 1D structure for organisms is easy via modern sequencing
- Characterizing 3D molecular conformations is still comparatively hard
- Understanding the 2D secondary structures (base pairings) remains a crucial component of ribonomics research
- RNA folding prediction programs generate 2D structures given the 1D base sequence

GTFoldPython software project

GTFold – Overview I

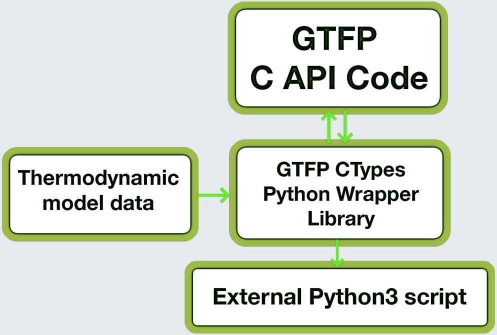
- Accurate and efficient RNA secondary structure prediction is an important open problem in computational molecular biology
- **Input:** The base sequence (FASTA) of the organism
  Output: MFE or suboptimal secondary structures (base pair data)
- GTFold is the first implementation of RNA secondary structure prediction by thermodynamic optimization for modern multi-core computers
- The speedup is useful to researchers working with very RNA sequences, such as RNA viral genomes

GTFold – Overview II

- Original GTFold produced command line only utilities
- Motivated by the need for a Python interface to get the same GTFold data
- Started work in the Fall of 2019 writing Python3 bindings for GTFold
- Hardest part of development work? Requirements for highly robust custom cross platform build scripts

GTFoldPython – Introduction

- GTFoldPython (GTFP): Python3 bindings based around the original GTFold sources written in C++
- Backend uses the Python3 C API
- Frontend interface is a wrapper library that uses CTypes to call the C API functions



GTFoldPython – Comparison – Python C API code

```
PyObject * GetMFEStructure(const char *baseSeq, ConsListCType_t consList, int consLength) {
    /* Error checking omitted ... */
    MFEStructRuntimeArgs_t rtArgs;
    InitMFEStructRuntimeArgs(&rtArgs);
    rtArgs.baseSeq = baseSeq;
    SetRTArgsSequenceLength(rtArgs, strlen(baseSeq));
    if(ParseGetMFEStructureArgs(consList, consLength, &rtArgs) != GTFPYTHON_ERRNO_OK) {
        FreeMFEStructRuntimeArgs(&rtArgs);
        return ReturnPythonNone();
    }
    if(InitGTFoldMFEStructureData(&rtArgs) != GTFPYTHON_ERRNO_OK) {
        FreeMFEStructRuntimeArgs(&rtArgs);
        return ReturnPythonNone();
    }
    double mfe = ComputeMFEStructure(&rtArgs);
    if(GetLastErrorCode() != GTFPYTHON_ERRNO_OK) {
        return ReturnPythonNone();
    }
    if(WRITEAUXFILES) {
        ConfigureOutputFileSettings();
        save_ct_file(outputFile, baseSeq, mfe);
    }
    char *dbMFEStruct = ComputeDOTStructureResult(rtArgs.numBases);
    PyObject *mfeTupleRes = PrepareMFETupleResult(mfe, dbMFEStruct);
    Free(dbMFEStruct);
    FreeMFEStructRuntimeArgs(&rtArgs);
    FreeGTFoldMFEStructureData(rtArgs.numBases);
    if(mfeTupleRes == NULL) {
        return ReturnPythonNone();
    }
    return mfeTupleRes;
```

GTFoldPython – Comparison – Wrapper library code

```python
## Library initialization code:
if GTFPConfig.PLATFORM_DARWIN:
    GTFoldPython._libGTFoldHandle = ctypes.cdll.LoadLibrary("GTFoldPython.dylib")
else:
    GTFoldPython._libGTFoldHandle = ctypes.PyDLL("GTFoldPython.so",
                                    mode=ctypes.RTLD_GLOBAL, use_errno=True)
@staticmethod
def _WrapCTypesFunction(funcname, restype=None, argtypes=None):
    return GTFoldPython._libGTFoldHandle.__getattr__(funcname)

@staticmethod
def GetMFEStructure(baseSeq, consList = []):
    """Get the MFE and MFE structure (in DOTBracket structure notation)
    :param baseSeq: A string of valid bases (ATGU/X)
    :param consList: A list of constraints on the MFE structure
    :return: A tuple (MFE as double, MFE structure as string in DOTBracket notation)
    :rtype: tuple
    """
    GTFoldPython._ConstructLibGTFold()
    resType = ctypes.py_object
    argTypes = [ GTFPTypes.CStringType,
                 GTFPTypes.FPConstraintsListType(consList),
                 ctypes.c_int ]
    libGTFoldFunc = GTFoldPython._WrapCTypesFunction("GetMFEStructure", resType, argTypes)
    (mfe, mfeStruct) = libGTFoldFunc(GTFPTypes.CString(baseSeq),
                                     GTFPTypes.FPConstraintsList(consList),
                                     len(consList))
    return (float(mfe), str(mfeStruct))
```

GTFoldPython – Example – Find MFE and MFE structure

External Python3 script source:

```python
import sys, os
from GTFoldPythonImportAll import *

GTFP.Init()
GTFP.Config(quiet = False, debugging = False, verbose = False, stdmsgout = "stderr")

baseSeqFPCons = "GCAUUGGAGAUGGCAUUCCUCCAUUAACAAACCGCUGCGCCCGUAGCAGCUGAUGAUGCCUACAGA"
consListFP = GTFPUtils.ReadFPConstraintsFromFile("../Testing/TestData/tRNA/yeast.fa.cons")

(mfe, mfeDOTStruct) = GTFP.GetMFEStructure(baseSeqFPCons, consListFP)
print("MFE␣%1.3f␣=>␣MFE␣DOT␣STRUCT␣\"%s\"\n\n" % (mfe, mfeDOTStruct))
```

Terminal output printed upon invoking the script above:

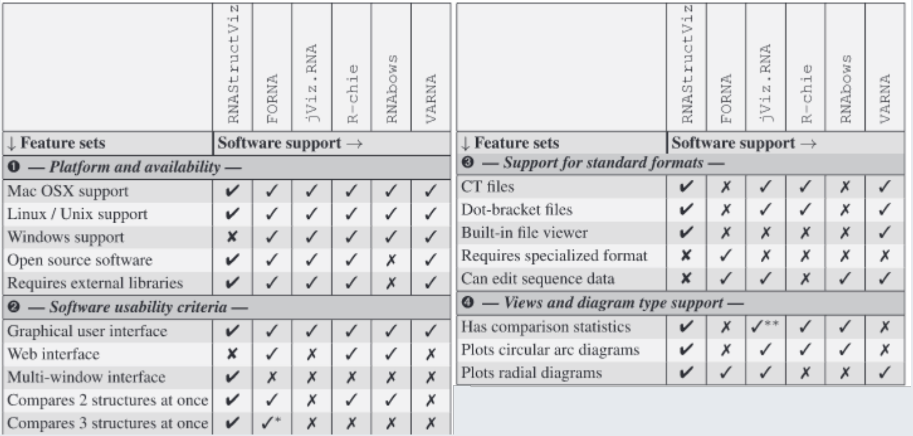
```
MFE -17.200 => MFE DOT STRUCT "(((((((((.........))))...........(((((.......))))).....)))))......
```

The RNAStructViz application

RNAStructViz: Graphical base pairing analysis

- RNAStructViz was a project developed by Professor Christine Heitsch and Dr. S. Cheney
- My first project working with the gtDMMB group in the Summer of 2018
    - Modernize the C++ source
    - Add support for enhanced graphics using the `cairo` library
    - Re-write the dated build scripts
    - Improve and support the project in the long term
- Key feature of RNAStructViz: Visualization and comparisons via arc diagrams of RNA secondary structures

RNAStructViz – Comparison of features

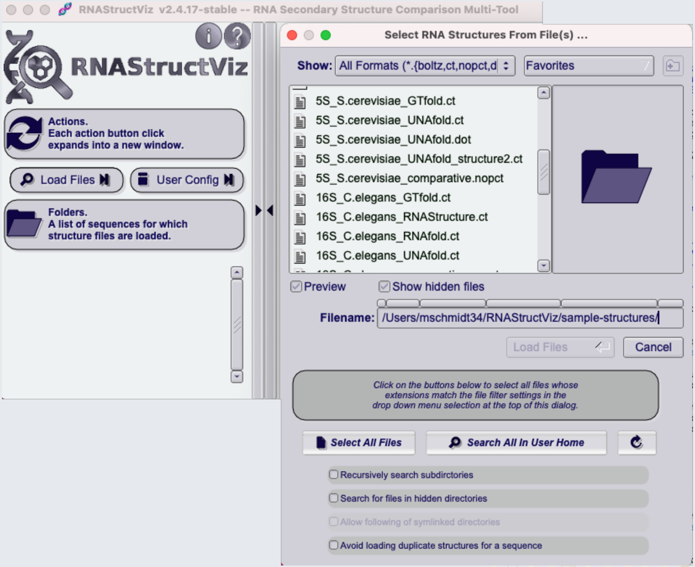


	RNAStructViz	FORNA	jViz.RNA	R-chie	RNAbows	VARNA
↓ **Feature sets**	**Software support →**					
❶ — ***Platform and availability*** —						
Mac OSX support	✔	✓	✓	✓	✓	✓
Linux / Unix support	✔	✓	✓	✓	✓	✓
Windows support	✘	✓	✓	✓	✓	✓
Open source software	✔	✓	✓	✓	✗	✓
Requires external libraries	✔	✓	✓	✓	✗	✓
❷ — ***Software usability criteria*** —						
Graphical user interface	✔	✓	✓	✓	✓	✓
Web interface	✘	✓	✗	✓	✓	✗
Multi-window interface	✔	✗	✗	✗	✗	✗
Compares 2 structures at once	✔	✓	✗	✓	✓	✗
Compares 3 structures at once	✔	✓*	✗	✗	✗	✗

	RNAStructViz	FORNA	jViz.RNA	R-chie	RNAbows	VARNA
↓ **Feature sets**	**Software support →**					
❸ — ***Support for standard formats*** —						
CT files	✔	✗	✓	✓	✗	✓
Dot-bracket files	✔	✗	✓	✓	✗	✓
Built-in file viewer	✔	✗	✗	✗	✗	✓
Requires specialized format	✘	✓	✗	✗	✗	✗
Can edit sequence data	✘	✓	✓	✗	✓	✓
❹ — ***Views and diagram type support*** —						
Has comparison statistics	✔	✗	✓**	✓	✓	✗
Plots circular arc diagrams	✔	✗	✓	✓	✓	✗
Plots radial diagrams	✔	✓	✓	✗	✗	✓

A comparison of selected features across related tools; an extended survey appears in the RNAStructViz WIKI.

RNAStructViz Screenshot – Loading sample structures I



RNAStructViz v2.4.17-stable -- RNA Secondary Structure Comparison Multi-Tool

Actions. Each action button click expands into a new window.

Folders. A list of sequences for which structure files are loaded.

Select RNA Structures From File(s) ...

Show: All Formats (*.{boltz,ct,nopct,d

- 5S_S.cerevisiae_GTfold.ct
- 5S_S.cerevisiae_UNAfold.ct
- 5S_S.cerevisiae_UNAfold.dot
- 5S_S.cerevisiae_UNAfold_structure2.ct
- 5S_S.cerevisiae_comparative.nopct
- 16S_C.elegans_GTfold.ct
- 16S_C.elegans_RNAStructure.ct
- 16S_C.elegans_RNAfold.ct
- 16S_C.elegans_UNAfold.ct

☑ Preview ☑ Show hidden files

Filename: /Users/mschmidt34/RNAStructViz/sample-structures/

Click on the buttons below to select all files whose extensions match the file filter settings in the drop down menu selection at the top of this dialog.

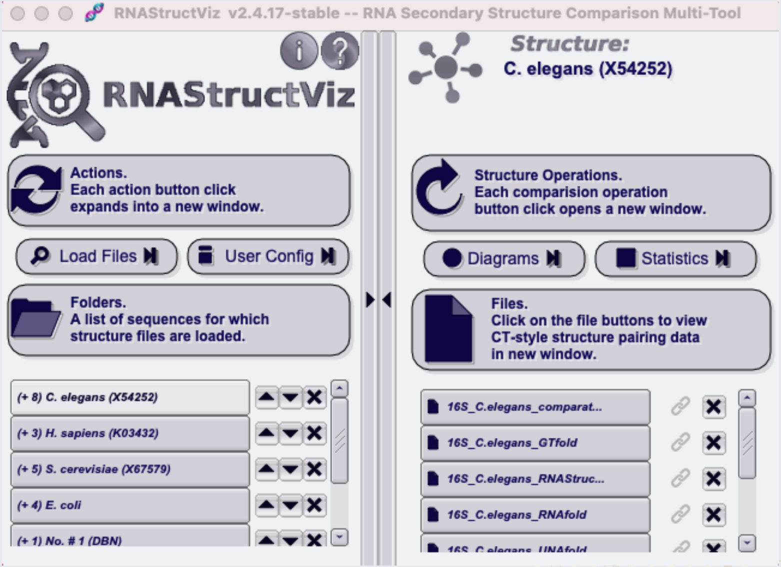
☐ Recursively search subdirctories

☐ Search for files in hidden directories

☐ Allow following of symlinked directories

☐ Avoid loading duplicate structures for a sequence

RNAStructViz Screenshot – Loading sample structures II



RNAStructViz v2.4.17-stable -- RNA Secondary Structure Comparison Multi-Tool

Structure:
C. elegans (X54252)

Actions.
Each action button click expands into a new window.

Structure Operations.
Each comparision operation button click opens a new window.

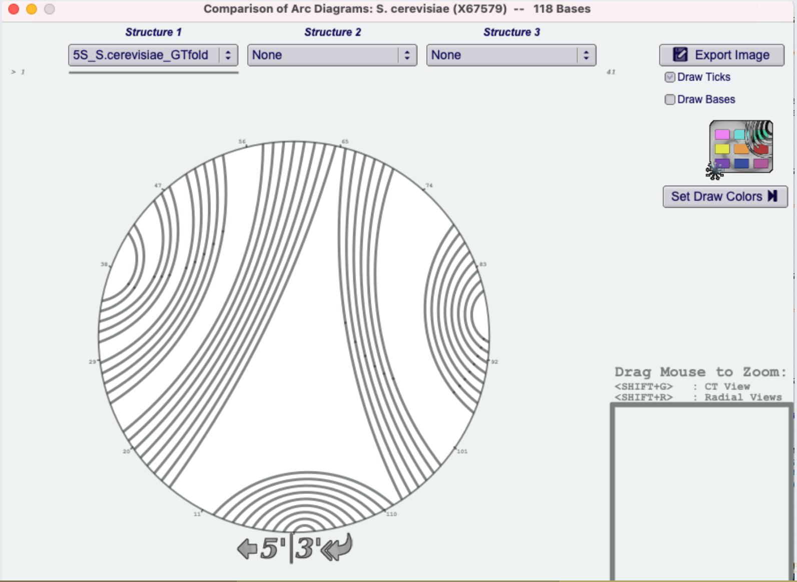
Folders.
A list of sequences for which structure files are loaded.

Files.
Click on the file buttons to view CT-style structure pairing data in new window.

(+ 8) C. elegans (X54252)
(+ 3) H. sapiens (K03432)
(+ 5) S. cerevisiae (X67579)
(+ 4) E. coli
(+ 1) No. # 1 (DBN)

16S_C.elegans_comparat...
16S_C.elegans_GTfold
16S_C.elegans_RNAStruc...
16S_C.elegans_RNAfold
16S_C.elegans_UNAfold

RNAStructViz Screenshot – Arc diagram window



Comparison of Arc Diagrams: S. cerevisiae (X67579) -- 118 Bases

Structure 1: 5S_S.cerevisiae_GTfold

Structure 2: None

Structure 3: None

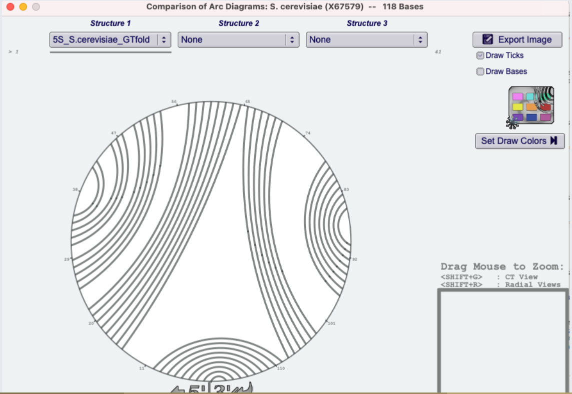
☑ Draw Ticks

□ Draw Bases

```
Drag Mouse to Zoom:
<SHIFT+G>    : CT View
<SHIFT+R>    : Radial Views
```

Arc diagram window – Discussion

- The bases from position #1 to #LengthOfBaseSequenceString at equidistant spacings around a circle
- The sequentially numbered base pairs are ordered around the circle counter-clockwise starting from the bottom
- An arc connecting paired bases is drawn within the circle



Comparison of Arc Diagrams: S. cerevisiae (X67579) -- 118 Bases

Structure 1: 5S_S.cerevisiae_GTfold

Structure 2: None

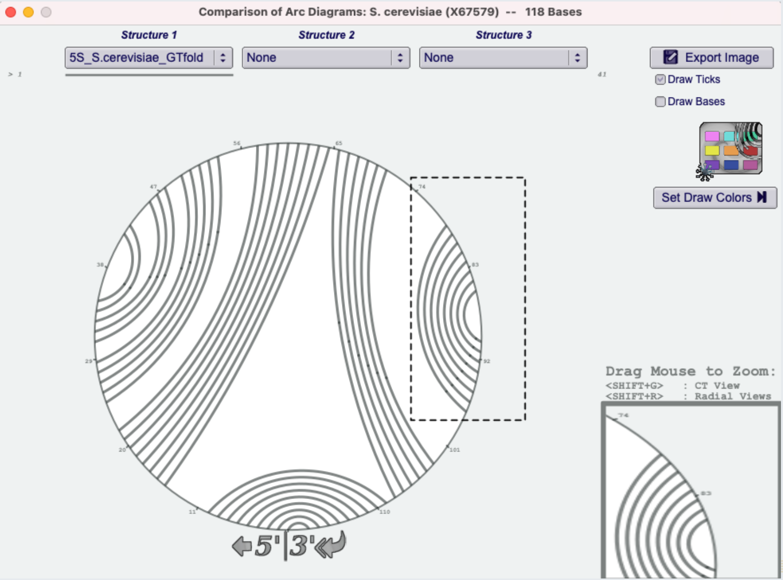
Structure 3: None

☑ Draw Ticks

□ Draw Bases

Drag Mouse to Zoom:
```
<SHIFT+G>   : CT View
<SHIFT+R>   : Radial Views
```

Arc diagram window – Zoom select



Comparison of Arc Diagrams: S. cerevisiae (X67579) -- 118 Bases

Structure 1: 5S_S.cerevisiae_GTfold

Structure 2: None

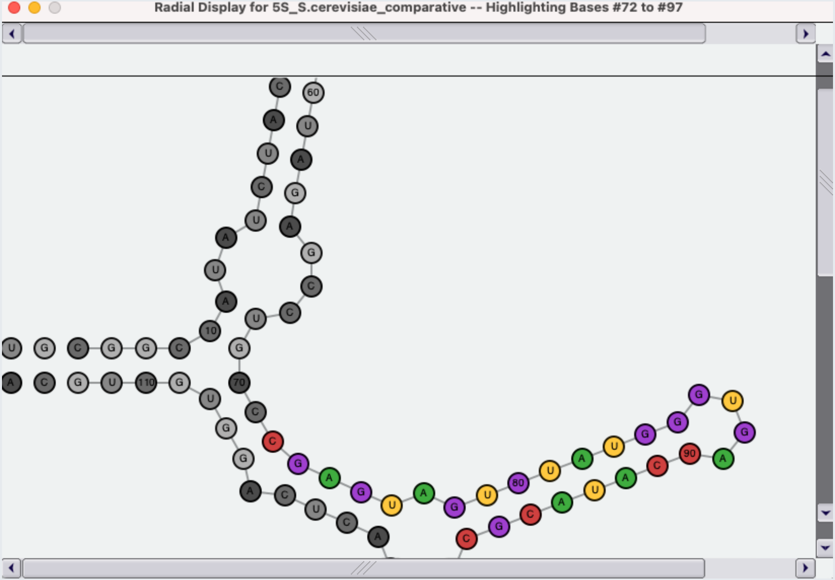
Structure 3: None

☑ Draw Ticks

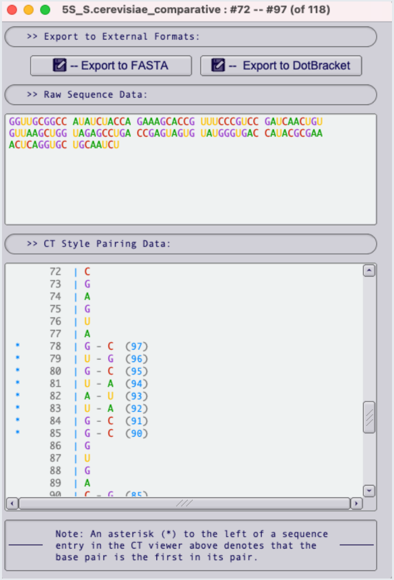
□ Draw Bases

```
Drag Mouse to Zoom:
<SHIFT+G>    : CT View
<SHIFT+R>    : Radial Views
```

Arc diagram zoom – Radial layout visualization


Radial Display for 5S_S.cerevisiae_comparative -- Highlighting Bases #72 to #97

Arc diagram zoom – CT segment visualization



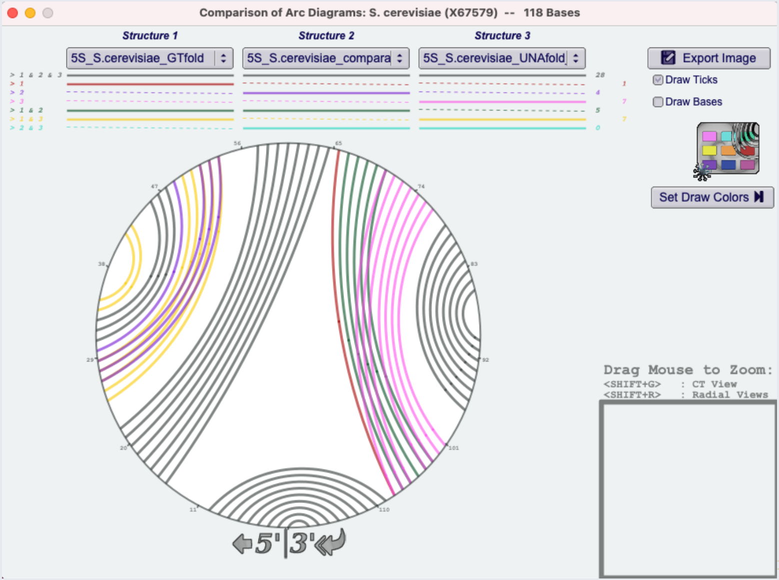
5S_S.cerevisiae_comparative : #72 -- #97 (of 118)

```
GGUUGCGGCC AUAUCUACCA GAAAGCACCG UUUCCCGUCC GAUCAACUGU
GUUAAGCUGG UAGAGCCUGA CCGAGUAGUG UAUGGGUGAC CAUACGCGAA
ACUCAGGUGC UGCAAUCU
```

```
     72  | C
     73  | G
     74  | A
     75  | G
     76  | U
     77  | A
*    78  | G - C  (97)
*    79  | U - G  (96)
*    80  | G - C  (95)
*    81  | U - A  (94)
*    82  | A - U  (93)
*    83  | U - A  (92)
*    84  | G - C  (91)
*    85  | G - C  (90)
     86  | G
     87  | U
     88  | G
     89  | A
     90  | C - G  (85)
```

Note: An asterisk (*) to the left of a sequence entry in the CT viewer above denotes that the base pair is the first in its pair.

Arc diagram window – Comparing multiple structures



Comparison of Arc Diagrams: S. cerevisiae (X67579) -- 118 Bases

Structure 1: 5S_S.cerevisiae_GTfold

Structure 2: 5S_S.cerevisiae_compara

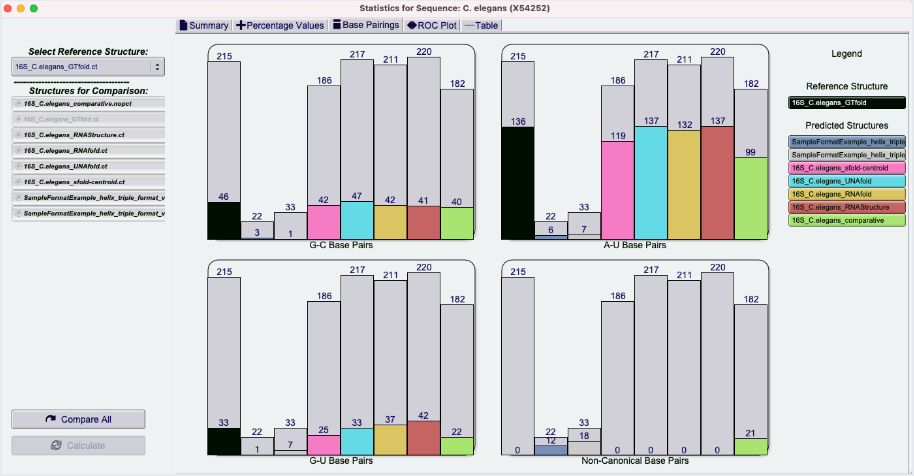
Structure 3: 5S_S.cerevisiae_UNAfold

☑ Draw Ticks

☐ Draw Bases

Drag Mouse to Zoom:
<SHIFT+G> : CT View
<SHIFT+R> : Radial Views

RNAStructViz Screenshot – Statistics window



Statistics for Sequence: C. elegans (X54252)

Select Reference Structure: 16S_C.elegans_GTfold.ct

Structures for Comparison:

- 16S_C.elegans_comparative.nopct
- 16S_C.elegans_GTfold.ct
- 16S_C.elegans_RNAStructure.ct
- 16S_C.elegans_RNAfold.ct
- 16S_C.elegans_UNAfold.ct
- 16S_C.elegans_sfold-centroid.ct
- SampleFormatExample_helix_triple_format_v
- SampleFormatExample_helix_triple_format_v

Legend

Reference Structure

- 16S_C.elegans_GTfold

Predicted Structures

- SampleFormatExample_helix_triple
- SampleFormatExample_helix_triple
- 16S_C.elegans_sfold-centroid
- 16S_C.elegans_UNAfold
- 16S_C.elegans_RNAfold
- 16S_C.elegans_RNAStructure
- 16S_C.elegans_comparative

Wrapping up

Summary of accomplishments with gtDMMB software I

- Success modernizing and enhancing the source code for these projects in computational and mathematical biology
- Success in modernizing and extending build scripts to support installation on MacOS, Linux and Unix-based systems
- A few of the software projects we worked on:
  - RNAStructViz
  - GTFold (CMake for MacOS and Linux)
  - GTFoldPython

Summary of accomplishments with gtDMMB software II

- Application note re-introducing our new work on RNAStructViz published in *Bioinformatics* in 2021
- Sister RNA labs that helped with testing and/or use our software include:
  - Computational RNA Genomics Lab at University of California Davis
  - Laederach Lab at the University of North Carolina at Chapel Hill
  - Mathews Lab at the University of Rochester

Concluding remarks

The End

Questions?

Comments?

Feedback?

Thank you for your time!